STA 602 Homework 6

Yutong Shao

2023-03-06

Problem Setting

Researchers are studying the length of life (lifetime) following a particular medical intervention, such as a new surgical treatment for heart disease, where the study consists of 12 patients. Specifically, the number of years before death for each is

$$3.4, 2.9, 1.2+, 1.4, 3.2, 1.8, 4.6, 1.7+, 2.0+, 1.4+, 2.8, 0.6+$$

where the + indicates that the patient was alive after x years, but the researchers lost contact with the patient after that point in time.

One way we can model this data is in the following way:

$$\begin{aligned} X_i &= \begin{cases} Z_i, & Z_i \leq c_i \\ c_i, & Z_i > c_i \end{cases} \\ Z_1, \ldots, Z_n | \theta &\overset{iid}{\sim} Gamma(r, \theta) \\ \theta &\sim Gamma(a, b) \end{aligned}$$

where a, b, and r are known. In addition, we know:

- c_i is the censoring time for patient i, which is fixed, but known only if censoring occurs.
- X_i is the observation
  - if the lifetime is less than c_i then we get to observe it ($X_i = Z_i$),
  - otherwise all we know is the lifetime is greater than c_i ($X_i = c_i$).
- θ is the parameter of interest—the rate parameter for the lifetime distribution.
- Z_i is the lifetime for patient i, however, this is not directly observed.

The probability density function (pdf) associated consists of two point masses:one at Z_i and one at c_i. The formula is

$$p(x_i|z_i) = \mathbb{1}(x_i = z_i)\mathbb{1}(z_i \leq c_i) + \mathbb{1}(x_i = c_i)\mathbb{1}(z_i > c_i).$$

Now we can easily find the full conditionals (derived in class and reproduced below). Notice that z_i is conditionally independent of z_j given θ for $i \neq j$. This implies that x_i is conditionally independent of x_j given z_i for $i \neq j$. Now we have

$$\begin{aligned} p(z_i|z_{-i}, x_{1:n}, \theta) &= p(z_i|x_i, \theta) \\ &\underset{z_i}{\propto} p(z_i, x_i, \theta) \\ &= p(\theta)p(z_i|\theta)p(x_i|z_i, \theta) \\ &\underset{z_i}{\propto} p(z_i|\theta)p(x_i|z_i, \theta) \\ &= p(z_i|\theta)p(x_i|z_i). \end{aligned}$$

There are now two cases to consider. If $x_i \neq c_i$, then $p(z_i|\theta)p(x_i|z_i)$ is only non-zero when $z_i = x_i$. The density devolves to a point mass at x_i. This corresponds to the case where z_i is observed, so x_i is the observed value and we should always sample this value. Practically speaking, we do not sample this value when running the Gibbs sampler.

If $x_i = c_i$, then the density becomes $p(x_i|z_i) = \mathbb{1}(z_i > c_i)$, so

$$p(z_i|\cdots) \propto p(z_i|\theta)\mathbb{1}(z_i > c_i),$$

which is a truncated Gamma.

For the Gibbs sampler, we will use the current value of θ to impute the censored data. We will sample from the truncated gamma using a modified version of the iverse CDF trick. For the censored values of Z_i we know c_i. If we know θ (which we will in a Gibbs' sampler), we know the distribution of $Z_i|\theta \sim Gamma(r, \theta)$. Let F be the CDF of this distribution. Suppose we truncate this distribution to (c, ∞). The new CDF is

$$P(Z_i < z) = \frac{F(z) - F(c)}{1 - F(c)}.$$

Therefore Y is a sample from the truncated Gamma, as desired.

In the actual code for the Gibbs' sampler we do not sample the observed values. We simply impute the censored values using the method above.

You will find code below (that is also taken from class) that will help you with the remainder of the problem.

1. (5 points) Write code to produce trace plots and running average plots for the censored values for 200 iterations. Do these diagnostic plots suggest that you have run the sampler long enough? Explain.
2. (5 points) Now run the chain for 10,000 iterations and update your diagnostic plots (traceplots and running average plots). Report your findings for both traceplots and the running average plots for θ and the censored values. Do these diagnostic plots suggest that you have run the sampler long enough? Explain.
3. (5 points) Give plots of the estimated density of $\theta \mid \cdots$ and $z_9 \mid \cdots$. Be sure to give brief explanations of your results and findings. (Present plots for 10,000 iterations).
4. (5 points) Finally, let's suppose that $r = 10, a = 1, b = 100$. Do the posterior densities in part (c) change for $\theta \mid \cdots$ and $z_9 \mid \cdots$? Do the associated posterior densities change when $r = 10, a = 100, b = 1$? Please provide plots and an explanation to back up your answer. (Use 10,000 iterations for the Gibbs sampler).

Answers

Question (a)

(a) (5 points) Write code to produce trace plots and running average plots for the censored values for 200 iterations. Do these diagnostic plots suggest that you have run the sampler long enough? Explain.

```r
knitr::opts_chunk$set(cache=FALSE)
library(xtable)

set.seed(2023)

# Samples from a truncated gamma with
# truncation (t, infty), shape a, and rate b
# Input: t,a,b
# Output: truncated Gamma(a,b)
sampleTrunGamma <- function(t, a, b){
  # This function samples from a truncated gamma with
  # truncation (t, infty), shape a, and rate b
  p0 <- pgamma(t, shape = a, rate = b)
  x <- runif(1, min = p0, max = 1)
  y <- qgamma(x, shape = a, rate = b)
  return(y)
}

# Gibbs sampler for censored data
# Inputs:
  # this function is a Gibbs sampler
  # z is the fully observe data
  # c is censored data
  # n.iter is number of iterations
  # init.theta and init.miss are initial values for sampler
  # r,a, and b are parameters
  # burnin is number of iterations to use as burnin
# Output: theta, z
sampleGibbs <- function(z, c, n.iter, init.theta, init.miss, r, a, b, burnin = 1){

  z.sum <- sum(z)
  m <- length(c)
  n <- length(z) + m
  miss.vals <- init.miss
  res <- matrix(NA, nrow = n.iter, ncol = 1 + m)
  for (i in 1:n.iter){
    var.sum <- z.sum + sum(miss.vals)
    theta <- rgamma(1, shape = a + n*r, rate = b + var.sum)
    miss.vals <- sapply(c, function(x) {sampleTrunGamma(x, r, theta)})
    res[i,] <- c(theta, miss.vals)
  }
  return(res[burnin:n.iter,])
}

# set parameter values
r <- 10
a <- 1
b <- 1
# input data
z <- c(3.4,2.9,1.4,3.2,1.8,4.6,2.8)
c <- c(1.2,1.7,2.0,1.4,0.6)
```

```r
n.iter <- 200
init.theta <- 1
init.missing <- rgamma(length(c), shape = r, rate = init.theta)
# run sampler
res <- sampleGibbs(z, c, n.iter, init.theta, init.missing, r, a, b)
```

In figure 1 and 2 we see traceplots for 200 iterations of the Gibbs sampler. It is difficult to tell whether or not the sampler has failed to converge, thus, we turn to running average plots.

```r
plot(1:n.iter, res[,1], pch = 16, cex = .35,
     xlab = "Iteration", ylab = expression(theta),
     main = expression(paste("Traceplot of ", theta)))
```

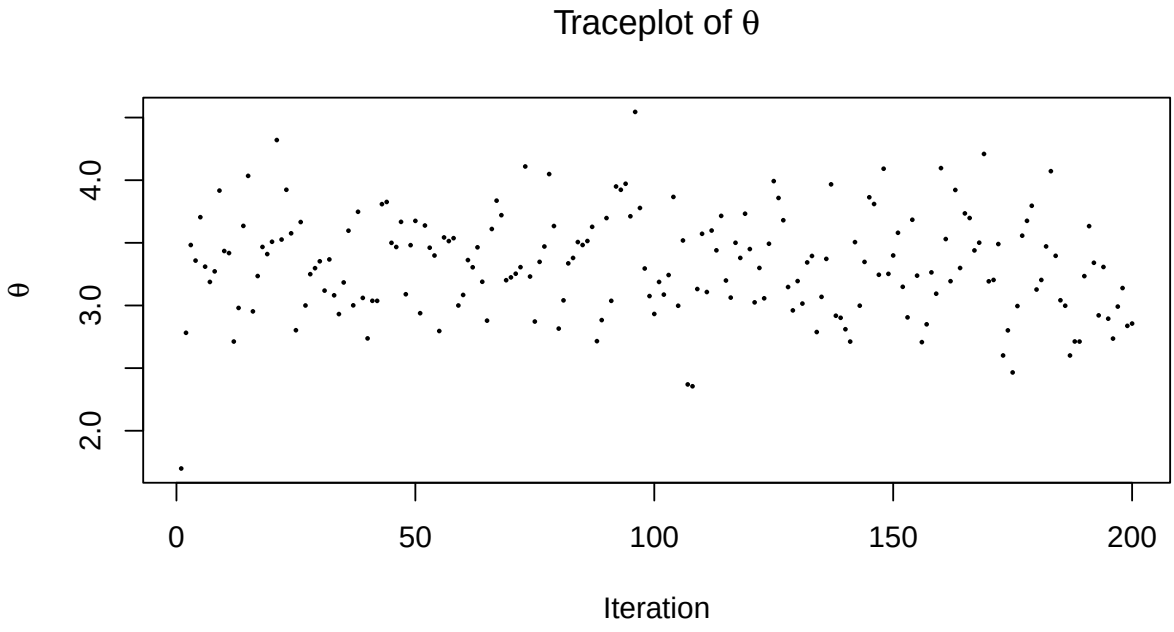

Figure 1: Traceplot of theta

In figures 3 and 4 we see running average plots for 200 iterations of the Gibbs sampler, where from all of these it is clear that after 200 iterations the sampler is having mixing issues, and **should be run for longer** to check that "it has not failed to converge."

```r
missing.index <- c(3,8,9,10,12)
par(mfrow=c(3,2))
for (ind in missing.index){
  x.lab <- bquote(z[.(ind)])
  plot(1:n.iter, res[,which(missing.index == ind)], pch = 16, cex = .35,
       xlab = "Iteration", ylab = x.lab,
       main = bquote(paste("Traceplot of ", .(x.lab))))
}
plot.new()
```

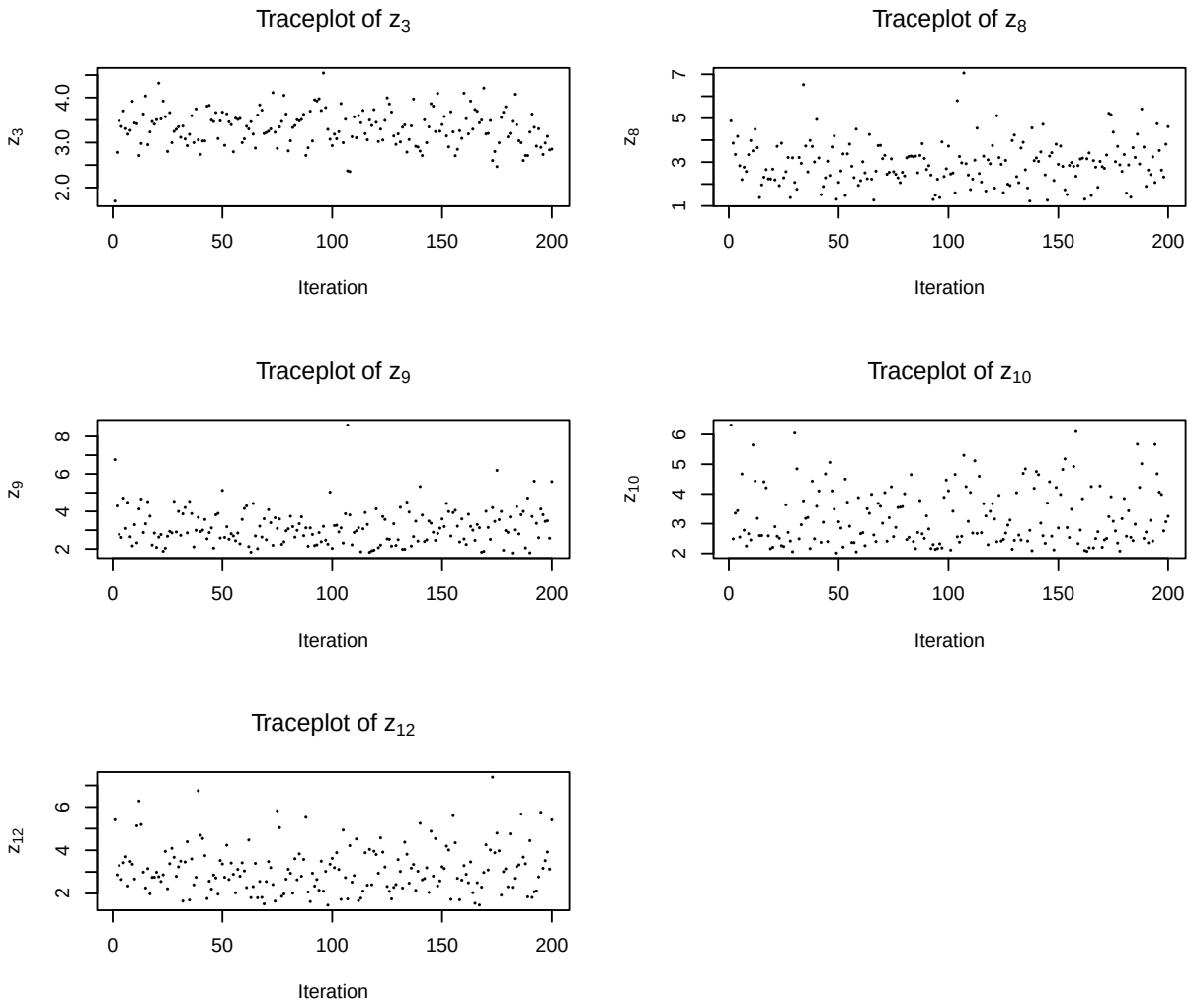

Figure 2: Traceplot of $z_3, z_8, z_9, z_{10}, z_{12}$.

```
# get running averages
run.avg <- apply(res, 2, cumsum)/(1:n.iter)
```

```
plot(1:n.iter, run.avg[,1], type = "l",
     xlab = "Iteration", ylab = expression(theta),
     main = expression(paste("Running Average Plot of ", theta)))
```

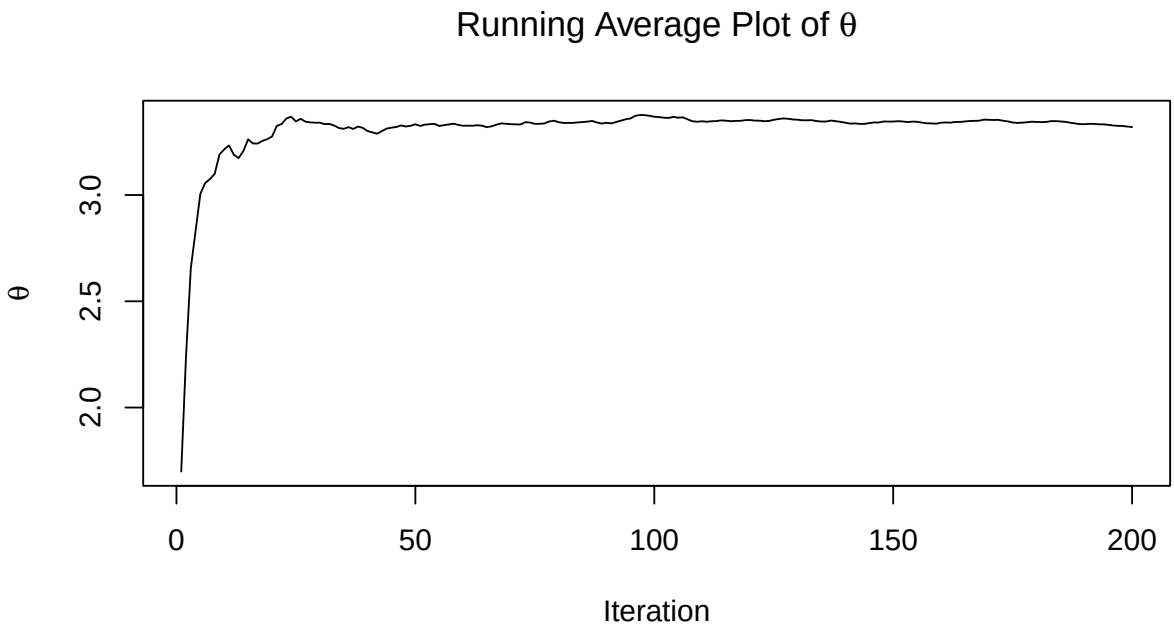

Figure 3: Running average plot of theta

```
par(mfrow=c(3,2))
missing.index <- c(3,8,9,10,12)
for (ind in missing.index){
  x.lab <- bquote(z[.(ind)])
  plot(1:n.iter, run.avg[,which(missing.index == ind)], type = "l",
       xlab = "Iteration", ylab = x.lab,
       main = bquote(paste("Running Average Plot of ", .(x.lab))))
}
plot.new()
```

Figures 5 and 6 do not provide meaningful inference at this point since the sampler **has not been run long enough**.

```
# calculate confidence interval
quantile(res[,1], c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 2.601021 4.091643
```

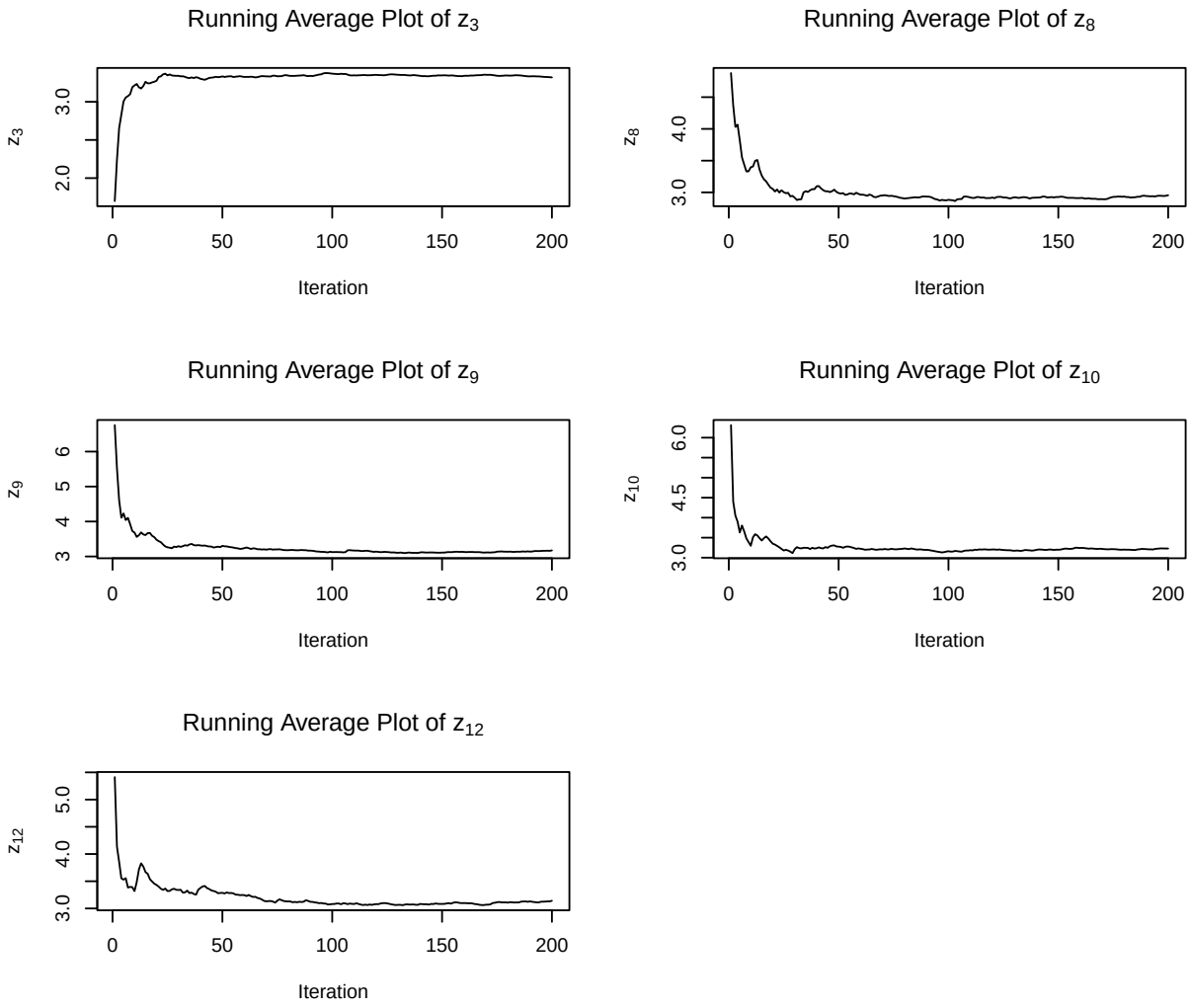

Figure 4: Running average plots of $z_3, z_8, z_9, z_{10}, z_{12}$.

```
# density plots

plot(density(res[,1]), xlab = expression(theta),
     main = expression(paste("Density of ", theta)))
abline(v = mean(res[,1]), col = "red")
abline(v = c(2.60, 4.09), col = "blue")
```

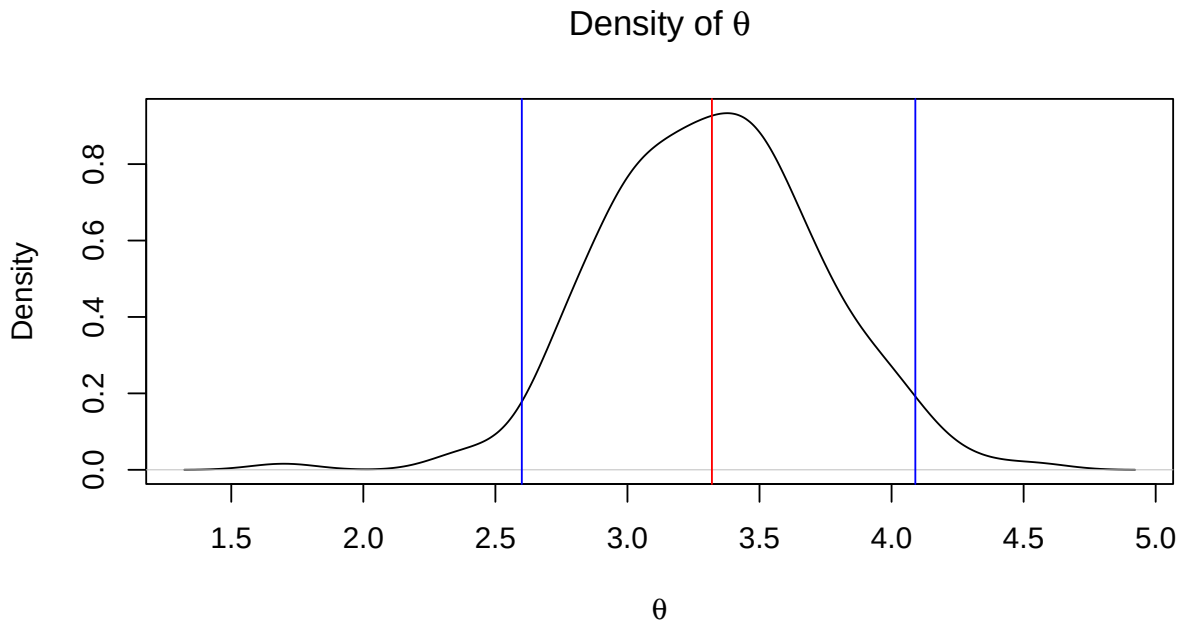

Figure 5: Estimated posterior density of theta

```
quantile(res[,4], c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 2.089185 5.646863
```

```
plot(density(res[,4]), xlab = expression(z[9]),
     main = expression(paste("Density of ", z[9])))
abline(v = mean(res[,4]), col = "red")
abline(v = c(2.08, 5.65), col = "blue")
```

Question (b)

(b) (5 points) Now run the chain for 10,000 iterations and update your diagnostic plots (traceplots and running average plots). Report your findings for both traceplots and the running average plots for θ and the censored values. Do these diagnostic plots suggest that you have run the sampler long enough? Explain.

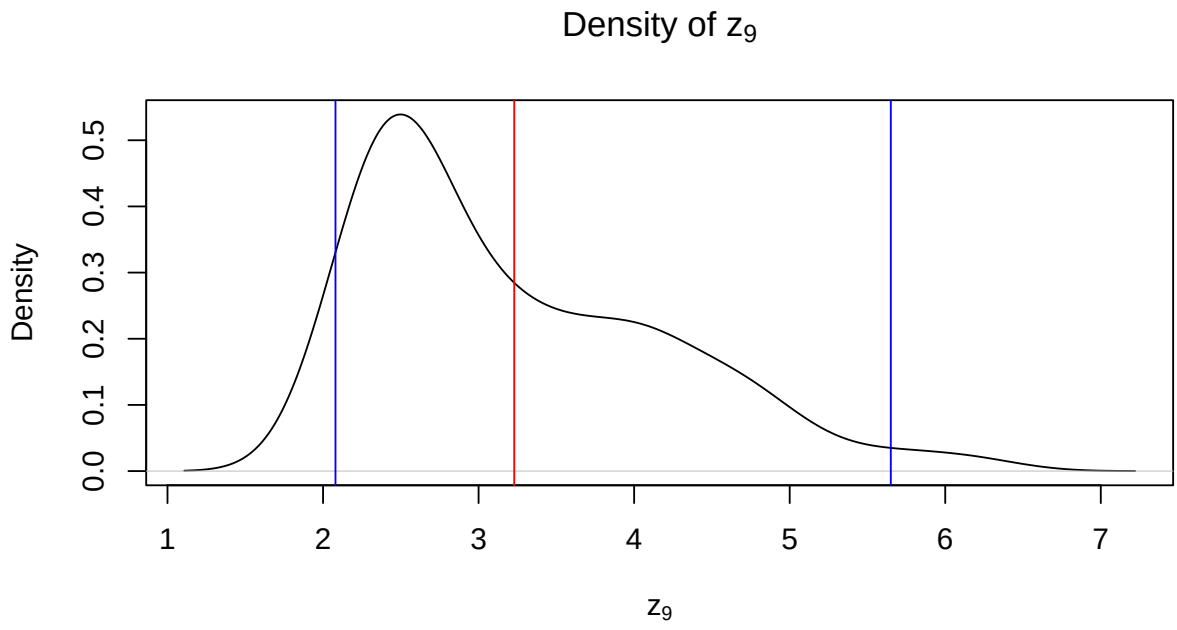

Figure 6: Estimated posterior density of z_9 (posterior mean in red).

```
n_iter_2 <- 10000
init.theta <- 1
init.missing <- rgamma(length(c), shape = r, rate = init.theta)
# run sampler
res_2 <- sampleGibbs(z, c, n_iter_2, init.theta, init.missing, r, a, b)
```

Traceplot of θ and censored values

```
plot(1:n_iter_2, res_2[,1], pch = 16, cex = .35,
     xlab = "Iteration", ylab = expression(theta),
     main = expression(paste("Traceplot of ", theta, " with 10000 iterations")))
```

```
missing.index <- c(3,8,9,10,12)
par(mfrow=c(3,2))
for (ind in missing.index){
  x.lab <- bquote(z[.(ind)])
  plot(1:n_iter_2, res_2[,which(missing.index == ind)], pch = 16, cex = .35,
       xlab = "Iteration", ylab = x.lab,
       main = bquote(paste("Traceplot of ", .(x.lab), " with 10000 iterations")))
}
plot.new()
```

Running Average plots of θ and censored values.

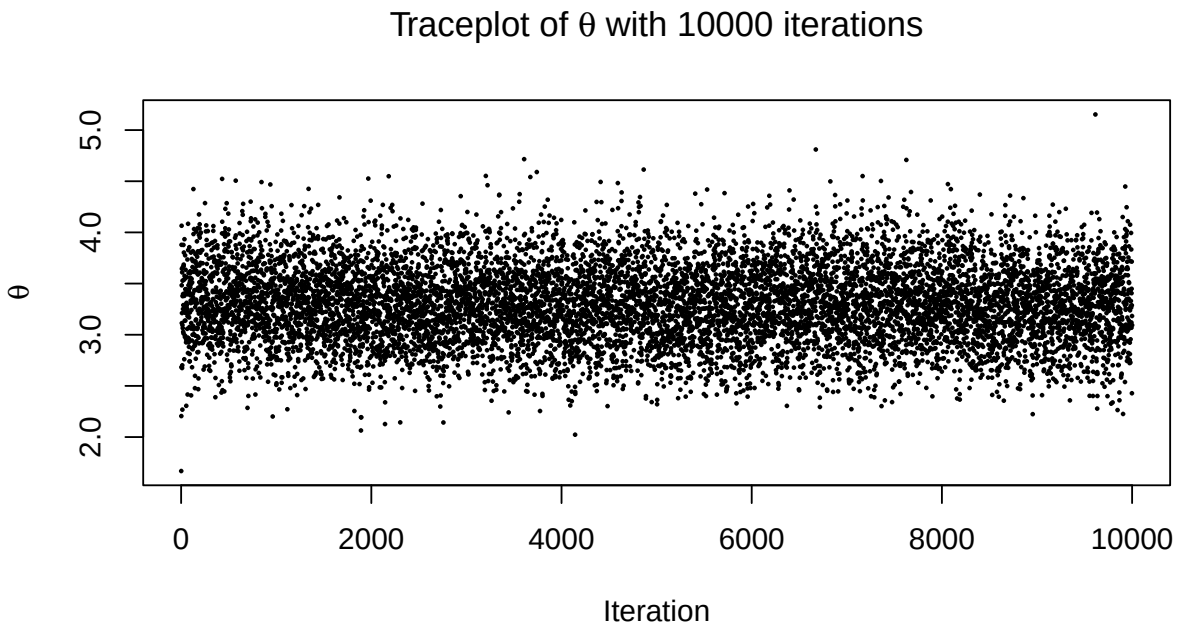

Figure 7: Traceplot of theta

```
# get running averages
run_avg_2 <- apply(res_2, 2, cumsum)/(1:n_iter_2)
```

```
plot(1:n_iter_2, run_avg_2[,1], type = "l",
     xlab = "Iteration", ylab = expression(theta),
     main = expression(paste("Running Average Plot of ", theta, " with 10000 iterations")))
```

```
par(mfrow=c(3,2))
missing.index <- c(3,8,9,10,12)
for (ind in missing.index){
  x.lab <- bquote(z[.(ind)])
  plot(1:n_iter_2, run_avg_2[,which(missing.index == ind)], type = "l",
       xlab = "Iteration", ylab = x.lab,
       main = bquote(paste("Running Average Plot of ", .(x.lab), " with 10000 iterations")))
}
plot.new()
```

From the Traceplots 7 and 8, it is not easy to infer whether they have converged or not. But the running average plots 9 and 10 clearly show that both θ and censored values have converged after 2000 iterations, because the lines tend to be flat as we keep iterating. Thus, we already run long enough for convergence.

Question (c)

(c) (5 points) Give plots of the estimated density of $\theta \mid \cdots$ and $z_9 \mid \cdots$. Be sure to give brief explanations of your results and findings. (Present plots for 10,000 iterations).

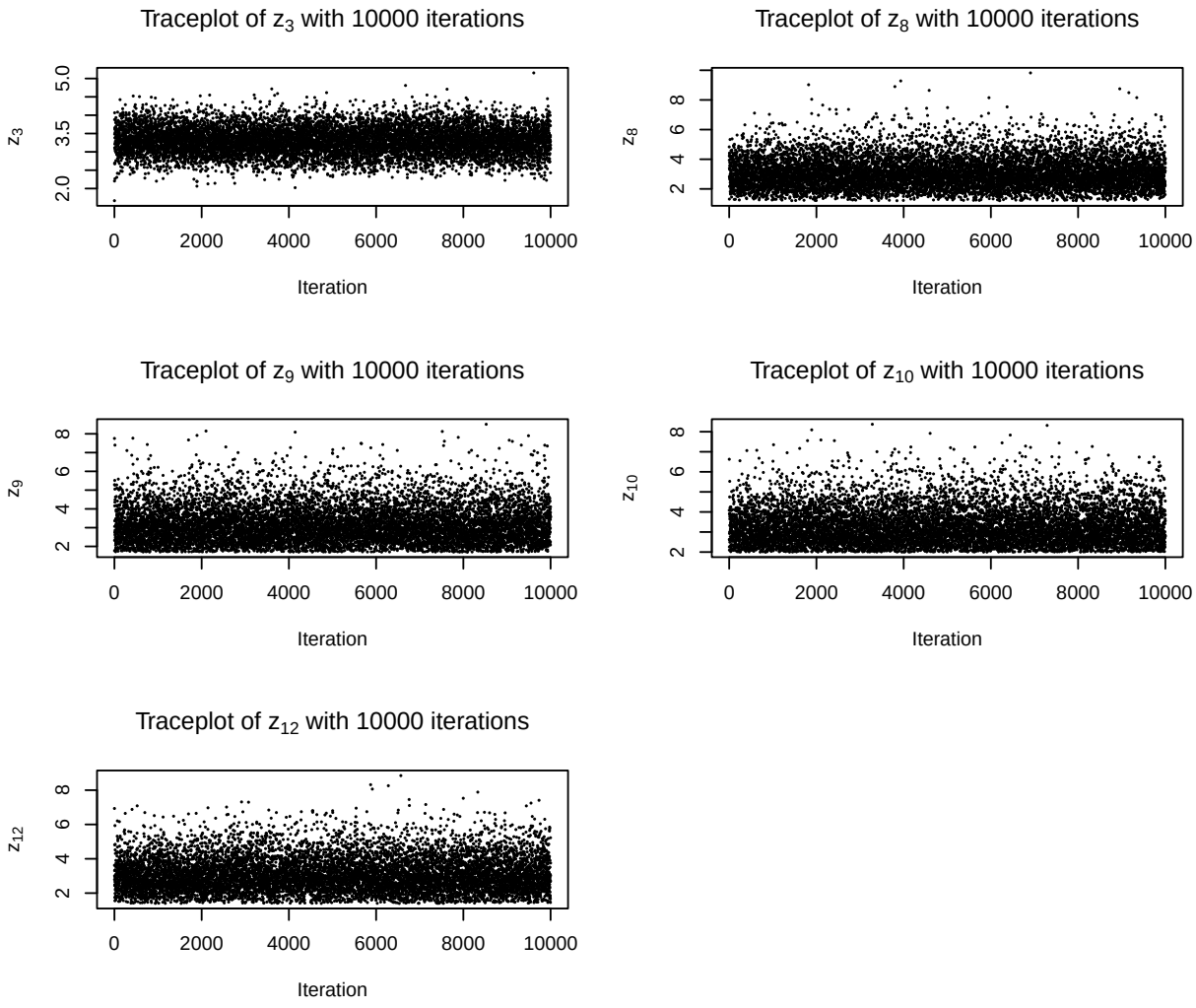

Figure 8: Traceplot of $z_3, z_8, z_9, z_{10}, z_{12}$.

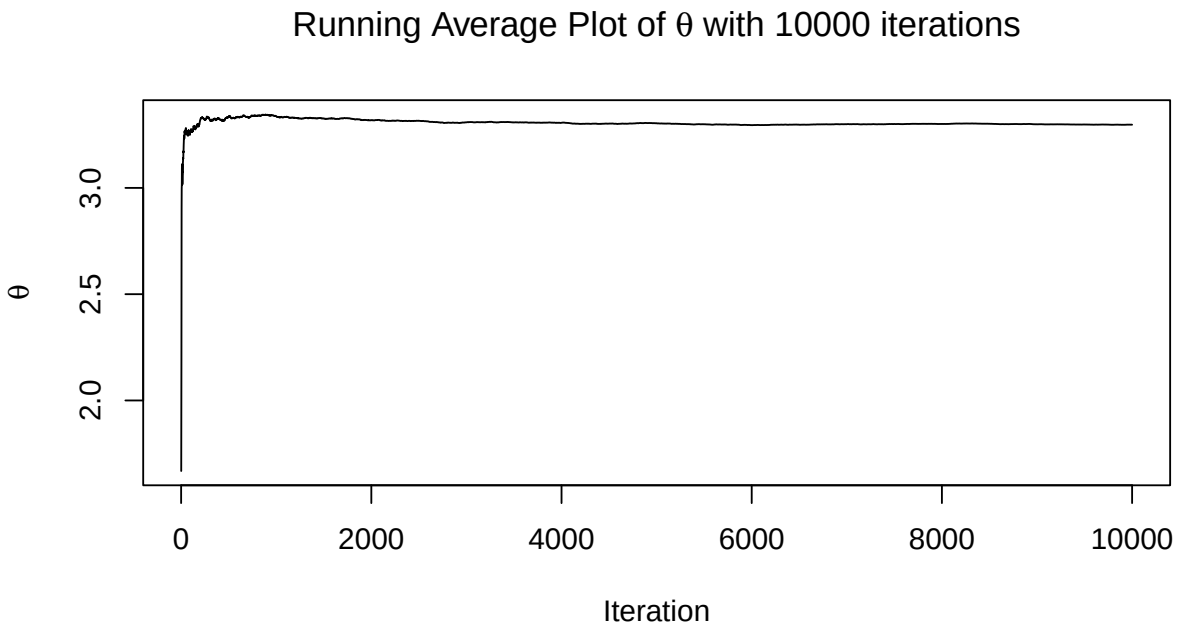

Figure 9: Running average plot of theta

```
# density plots
quantile(res_2[,1], c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 2.577513 4.046908
```

```
plot(density(res_2[,1]), xlab = expression(theta),
     main = expression(paste("Density of ", theta, " with 10000 iterations")))
abline(v = mean(res_2[,1]), col = "red")
abline(v = c(2.57, 4.05), col = "blue")
```

```
quantile(res_2[,4], c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 2.070974 5.526595
```

```
plot(density(res_2[,4]), xlab = expression(z[9]),
     main = expression(paste("Density of ", z[9])))
abline(v = mean(res_2[,4]), col = "red")
abline(v = c(2.07, 5.53), col = "blue")
```

The first graph shows the estimated density of θ is an updated Gamma distribution. This represents the approximated posterior density of an individual's lifetime before death. The posterior mean shifted slightly to the left as compared to the 200-iteration case.

The second graph is the estimated density of z_9, which is the imputed value of lifetime of individual 9. Here the density of z_9 is more skewed as the mode shifted more to left than the mean, comparing with the 200-iteration case. Thus, increasing the number of iterations is necessary to achieve reliable distributions.

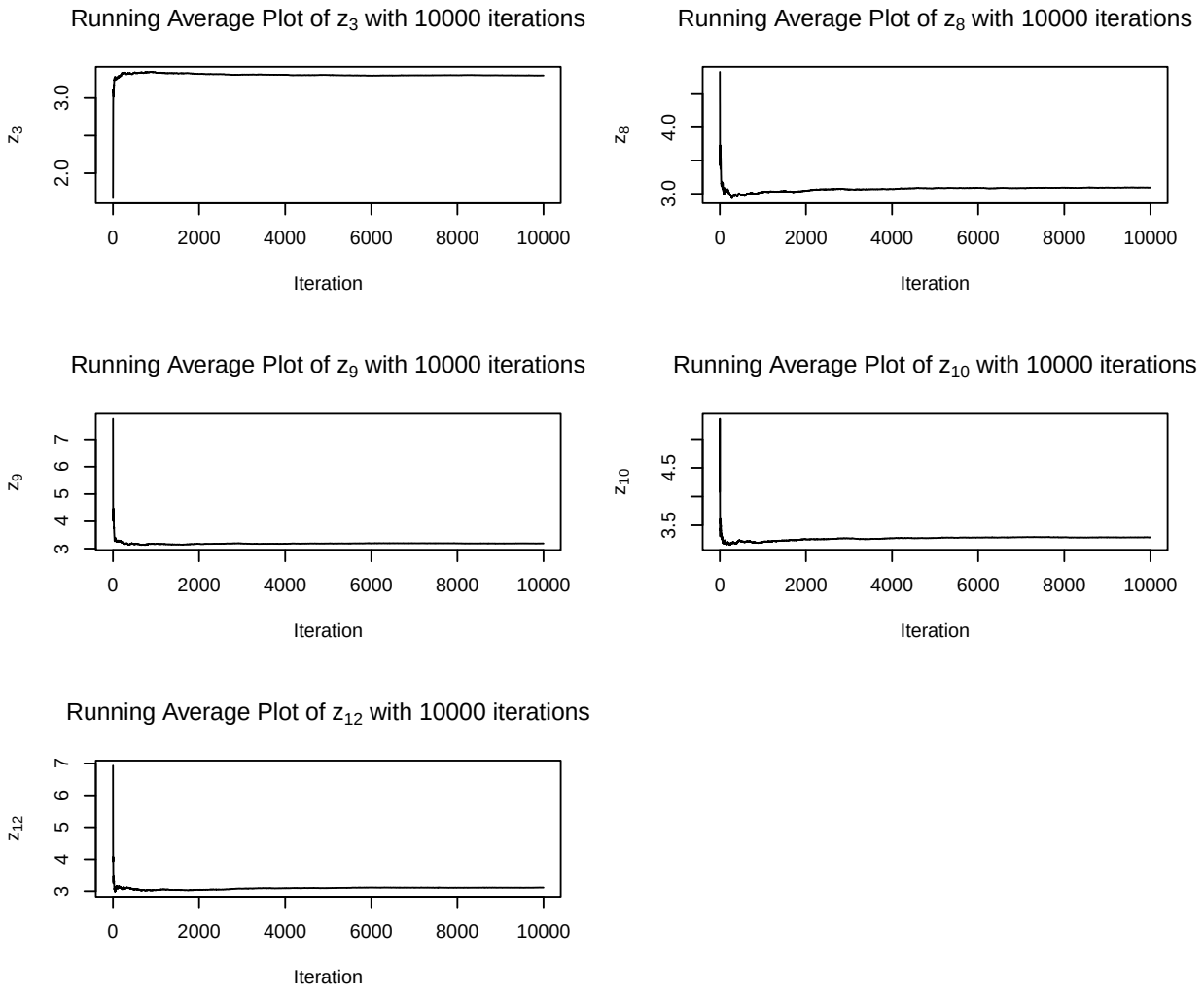

Figure 10: Running average plots of $z_3, z_8, z_9, z_{10}, z_{12}$.

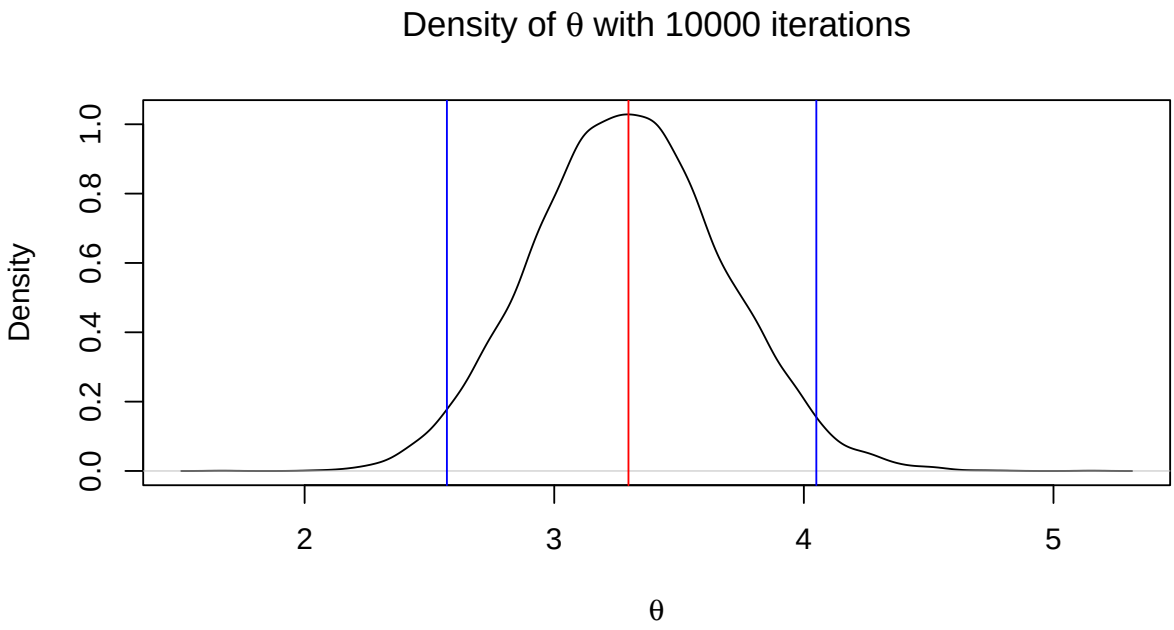

Figure 11: Estimated posterior density of theta

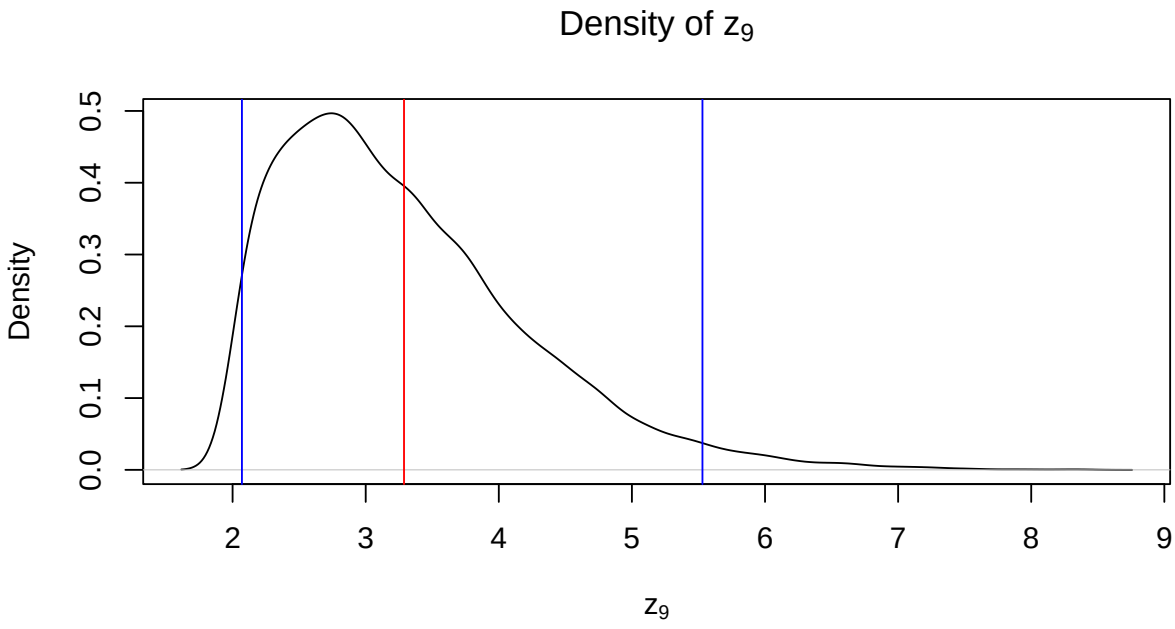

Figure 12: Estimated posterior density of z_9 (posterior mean in red).

Question (d)

(d) (5 points) Finally, let's suppose that $r = 10, a = 1, b = 100$. Do the posterior densities in part (c) change for $\theta \mid \cdots$ and $z_9 \mid \cdots$? Do the associated posterior densities change when $r = 10, a = 100, b = 1$? Please provide plots and an explanation to back up your answer. (Use 10,000 iterations for the Gibbs sampler).

$r = 10, a = 1, b = 100$

```
# new parameter values
r <- 10
a <- 1
b1 <- 100

# run sampler
res_3 <- sampleGibbs(z, c, n_iter_2, init.theta, init.missing, r, a, b1)
quantile(res_3[,1], c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.4629788 0.7311854
```

```
quantile(res_3[,4], c(0.025, 0.975))
```

```
##      2.5%     97.5%
##  7.952015 30.609114
```

```
# density plots

plot(density(res_3[,1]), xlab = expression(theta),
     main = expression(paste("Density of ", theta, " with 10000 iterations, b=100")))
abline(v = mean(res_3[,1]), col = "red")
abline(v = c(0.46, 0.73), col = "blue")
```

```
plot(density(res_3[,4]), xlab = expression(z[9]),
     main = expression(paste("Density of ", z[9], ", 10000 iterations, b=100")))
abline(v = mean(res_3[,4]), col = "red")
abline(v = c(8.06, 30.91), col = "blue")
```

$r = 10, a = 100, b = 1$

```
# new parameter values
r <- 10
a1 <- 100
b <- 1

# run sampler
res_4 <- sampleGibbs(z, c, n_iter_2, init.theta, init.missing, r, a1, b)
quantile(res_4[,1], c(0.025, 0.975))
```

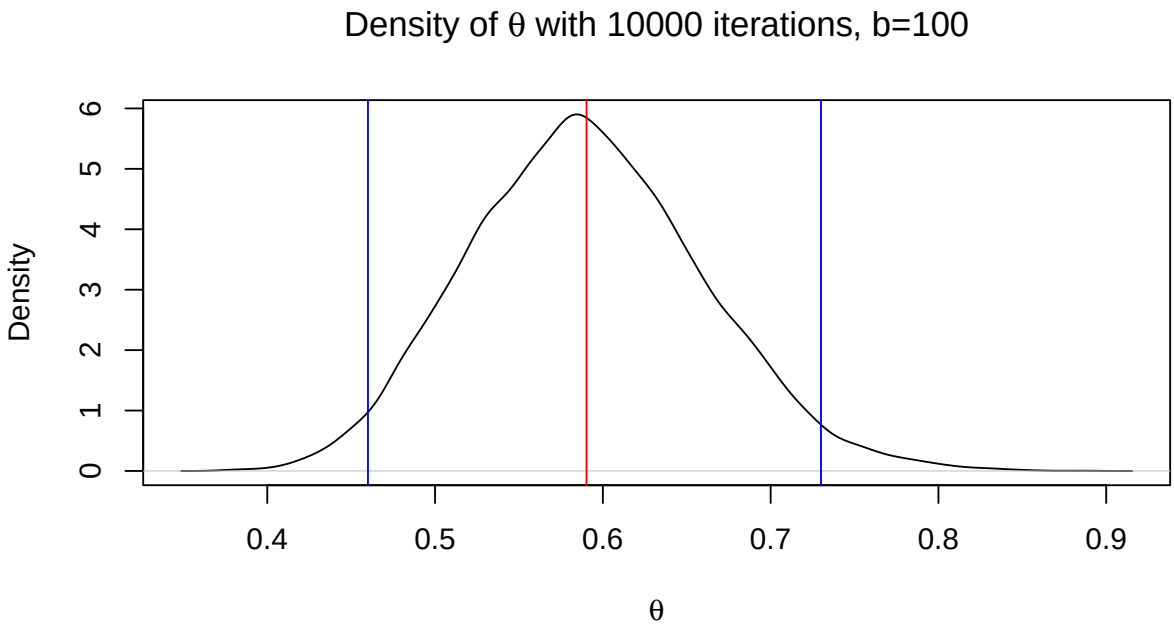

Figure 13: Estimated posterior density of theta

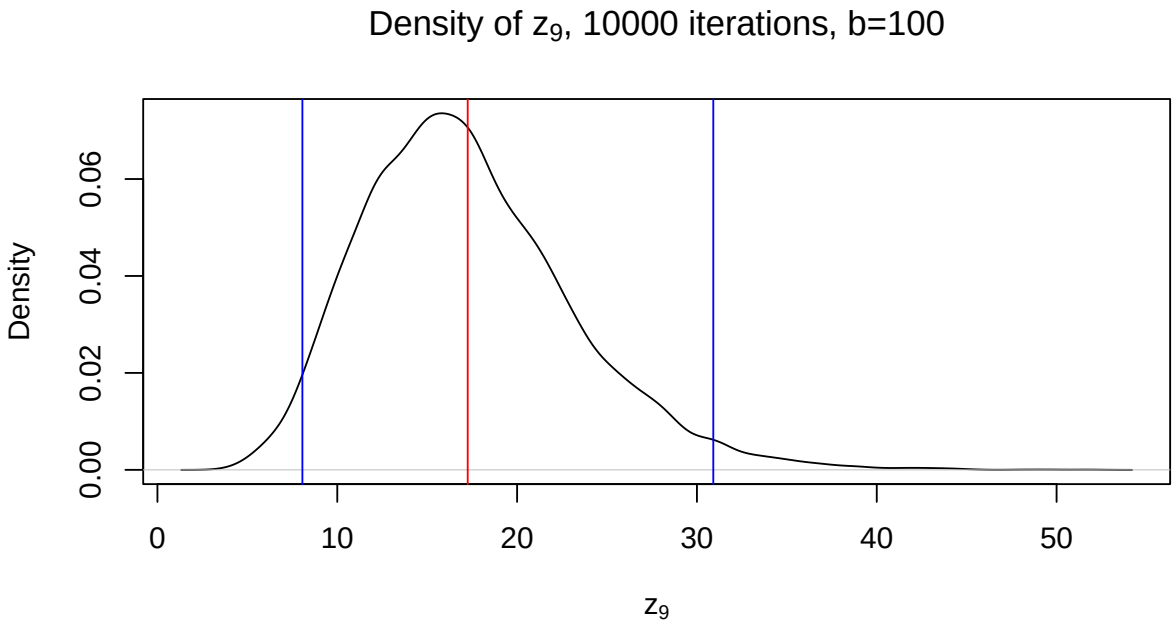

Figure 14: Estimated posterior density of z_9 (posterior mean in red).

```
##     2.5%    97.5%
## 6.304761 8.387699
```

```
quantile(res_4[,4], c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 2.008342 2.997592
```

```
# density plots
plot(density(res_4[,1]), xlab = expression(theta),
     main = expression(paste("Density of ", theta, ", 10000 iterations, a=100")))
abline(v = mean(res_4[,1]), col = "red")
abline(v = c(6.31, 8.34), col = "blue")
```

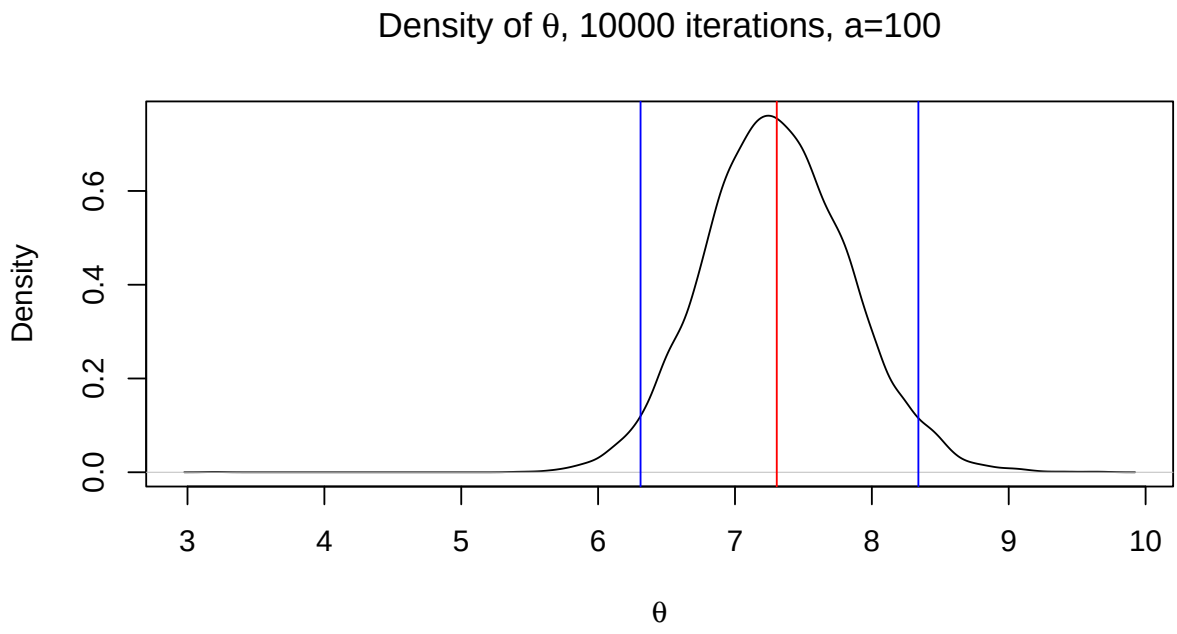

Figure 15: Estimated posterior density of theta

```
plot(density(res_4[,4]), xlab = expression(z[9]),
     main = expression(paste("Density of ", z[9], ", 10000 iterations, a=100")))
abline(v = mean(res_4[,4]), col = "red")
abline(v = c(2.01, 3.00), col = "blue")
```

The posterior density changes with different parameters, and the 95% confidence interval is very different.

- When $b = 100$ and others remains the same, the posterior values of θ decreases with most values concentrate on 0.6. While the posterior of z_9 increases a lot, with point mass around 15, which is not quite reasonable comparing other individual's lifetime.

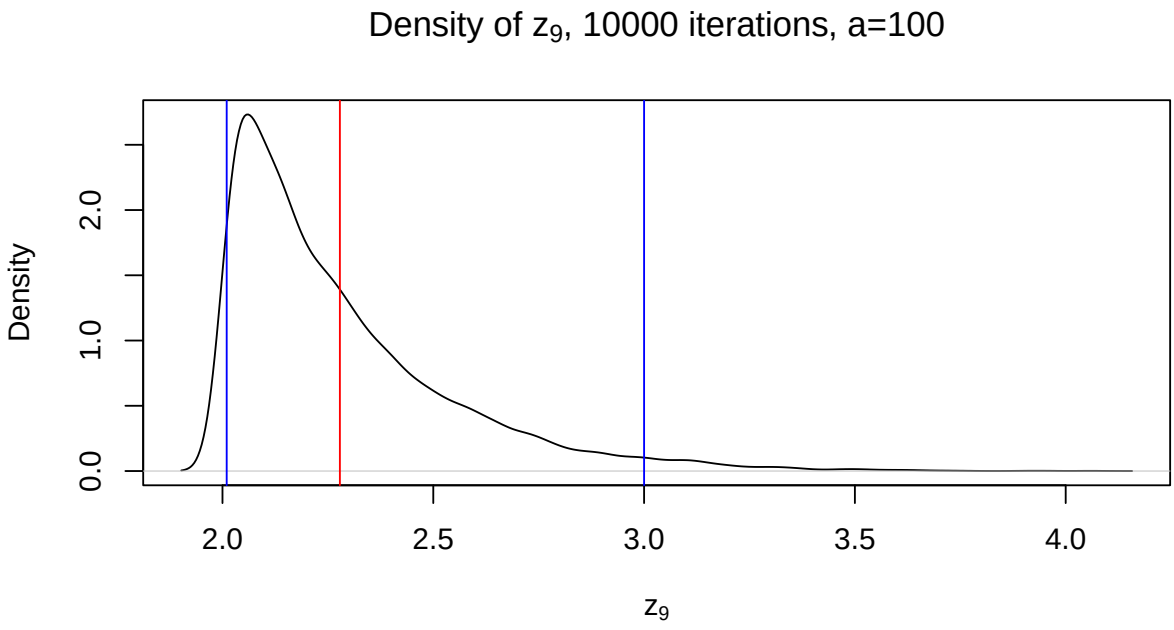

Figure 16: Estimated posterior density of z_9 (posterior mean in red).

- When $a = 100$ and others remains the same, the posterior values of θ increases with most values concentrate on around 7. While the posterior of z_9 slightly decreases, with point mass on around 2.1. And the density around $z_9 \in (0, 2)$ is very close to zero, meaning the lifetime of the 9th patient with censored data will probably have 2.2 to 2.5 years of living after treatment.

Thus, different choices of parameters will greatly influence the posterior density estimation. We should be careful on choosing the prior parameters.

References

Code was adapted from course materials provided by Prof. Rebecca C. Steorts.